DOCUMENT RESUME

ED 161 426 IR 006 348

AUTHOR Mello, James F.
TITLE Computerization of Synonymy Data from Biological Systematics. Museum Data Bank Research Report Number 9.
INSTITUTION Margaret Woodbury Strong Museum, Rochester, N.Y. Museum Data Bank Committee.
PUB DATE Sep 77
NOTE 21p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Automation; *Biological Sciences; *Classification; Definitions; Efficiency; *Electronic Data Processing; History; *Information Processing; Records (Forms); Taxonomy
IDENTIFIERS Morphology; Synonomy; *Systematics

ABSTRACT

An outline of biological systematics and the kind of data that comprise it is presented, together with a description of one data convention used in systematics and a method for computerizing synonymic and associated data. Systematics consist of descriptions of species, names applied to these species, and supportive field data such as geographic distribution. Historically, knowledge about plant and animal species is systematized through conventionalized descriptive terminology and the establishment of hierarchical relationships between species and other taxa. As the system grows, there is a need for greater precision in description and inclusion of more features among included species. Therefore, internal growth of systematic data is inevitable, and it creates management problems even for small groups of species. The computer can play a significant role in creating order. Three levels of description in systematics are: taxonomy--application of names to plants and animals; synonomy--recording of taxonomic history and referrals to that species; and morphologic description-detailing of distinguishing characteristics of a species. These levels are well standardized, thereby lending themselves to computerization. This study concentrates on developing efficient methods of data entry and computer processing for synonymic data, making these data increasingly useful as a source of information on research questions. (Author/EBR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED161426

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MUSEUM DATA BANK RESEARCH REPORT

Number 9

Computerization of synonymy data from biological systematics

by
James F. Mello

September, 1977

This report, or any portion thereof, may be reproduced for any non-commercial purpose provided appropriate credit is given to the author and the Museum Data Bank Committee as publisher.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. G. Chenhall

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Museum Data Bank Committee
Robert G. Chenhall, Chairman

Margaret Woodbury Strong Museum
700 Allen Creek Road
Rochester, N. Y.

IR006348

INTRODUCTION

Biological systematics is a broad field of scientific study. The long history of concentrated effort in systematics has resulted in a large volume of data, much of which is presented according to various complex but effective conventions. In this paper I try to present a feeling for the character of biological systematics by giving a brief outline of the field and of the kinds of data which comprise it. Following that I present a description of one type of data convention used in systematics (synonymy) and describe a method for computerizing synonymic and associated data.

For the reader who is not familiar with systematics the first part of the paper should provide a useful background. For the systematist the possible uses of the computer in handling synonymy and other kinds of data structures in systematics should be of chief interest.

Acknowledgements

The thoughts expressed in this paper were first presented orally to the Museum Data Bank Committee in March of 1975. An early draft of the paper was reviewed by Dr. Terry Erwin of the NMNH, and at a later stage by Dr. Gary Gautier, Cheif of the Automatic Data Processing Program at the NMNH. I am most grateful for the many useful suggestions and criticisms made by these readers, but of course I am solely responsible for the contents of the paper.

I owe a particular debt of gratitude to Dr. Jack Heller of the State University of New York at Stonybrook. He has worked very closely with me in the development of the concepts presented here, and has independently modified GRIPHOS Computer Programs to accommodate some of the more refractory data combinations found in synonymies.

The Data of Biological Systematics

Biological systematics is concerned with the biological elements of nature - the species of plants and animals - and includes naming them, describing their physical and behavioral characteristics, and recording their occurrence. Two basic objectives in systematics are to accurately record as much information as possible about species of animals and plants and to establish reliable criteria for discrimination among the species.

Organic species are invariably complex in form, the environments in which they live are multifaceted, and the interactions in which they are involved are numerous and complicated. In short, the data available to the systematist about any one species are far more than he can record. Therefore, in practice, the systematist must carefully study the organisms he works on in order to gain insight into their basic structural functional and behavioral characteristics. He then describes those <u>selected</u> features or actions which he regards as

characteristic for the species and uses them to discriminate this species from others. Far more data are available in nature about any species than can reasonably be recorded, and therefore systematists select what they consider to be diagnostic characteristics for use in description of their species, and for discrimination among them.

The natural world is populated by millions of species of animals and plants, each finely attuned to its environment. In order to bring the tremendous volume of information available in the natural world within the scope of the mind of man, it has been necessary to drastically narrow the scope of detailed biological descriptions, as described above, and to conventionalize the frame of reference within which these descriptions are presented.

This data selectivity and conventionalization has added a strong historical element to biological systematics. The establishment of certain morphologic features or behavioral characteristics of a species as diagnostic means that all workers on the group of organisms containing that species must know what the diagnostic characteristics are in order to recognize it. Therefore great care is taken in conserving the data of biological systematics. Another aspect to the historical nature of systematics is the value of specimen collections. Systematists recognize that the morphologic characters which they select as diagnostic for any organism may prove to be inadequate, and they realize that a great deal more information is contained in any specimen of plant or animal than can be described. Great effort is made to record observational information, such as geographic locality data, with specimens stored in systematic collections, because without such supporting data the value of the specimens is greatly diminished.

In summary, the data of systematics consist in part of diagnosis or descriptions of species, the names applied to these species, and supportive field data such as geographic distribution. In historical perspective, knowledge about animal and plant species is systematized through the use of conventionalized descriptive terminology and also through the establishment of hierarchical relationships between species and other taxa. The specimen collections also serve as an historical resource which can be called upon for new data, to check old interpretations, and to test new interpretaions.

Problems with Systematics Data

In the early days of biological systematics, when relatively few kinds of plants and animals were available for description, only a few features of morphology or behavior had to be described in order to discriminate one animal or plant from another. In addition, the things described were often part of man's everyday environment, and this familiarity reduced the need to provide detailed descriptions. For example, far fewer terms would suffice to describe a cow than a llama for the English farmer, and vice versa for the Peruvian peasant.

As the era of discovery in systematics progressed, and new groups were discovered by researchers, the tendency was to briefly describe the new species and other taxa as they were discovered because the limited number of taxa then known did not require more elaborate description for adequate characterization. However in many groups, as more taxa have been discovered and as more penetrating research has been focused on them, it has been necessary to develop more and more precise descriptions. This increase in the number of features described, coupled with frequent changes in the meanings of terms used in description, presents a serious problem to systematists. The progressive, historical character of systematics cannot adjust well to the increase in descriptive terms or to changes in their meaning. Adjustments, in the form of new monographic treatments of groups of species and higher taxa, are constantly being made; but, while these may serve as modern syntheses for the benefit of present workers, they also add a burden to systematics as an historical science because now the new as well as the old interpretations must be retained for posterity.

In summary, the need for greater and greater precision in description, and the inclusion of more and more features among those described, creates a kind of internal very complex growth in the body of systematics data which is added to the tremendous outward growth as more and more species are discovered. In a fashion these complexities are preserved for the future in the scientific literature, in specialized compendia created for some groups, and in various kinds of cross-reference files and listings prepared by individual systematists. This growth and shift of concepts in systematics can never be "managed" in an orderly fashion and efforts to completely organize the data of systematics, even for small groups of species, will be very difficult. Yet the historical perspective of systematics is important to the science, and efforts should be made to make these data as available as possible to modern workers. The computer can play a significant role in this work.

Use of the Computer in Systematics

The computer is being used more and more intensively in all phases of systematics studies. The chief reason for this may be that many systematists, and particularly younger scientists, have become more familiar with the computer and what it can do for them. Other contributing factors are the increasing availability of computers, the decreasing complexity of access to computers, the increasing sophistication and variety of statistical and other kinds of computer programs relevant to systematics, and the vast increase in the machine capabilities which have made possible the performance of jobs which were beyond the capabilities of earlier equipment.

The computer has been used most intensively in two ways in systematics. Individual scientists have used it extensively for statistical analyses and other data recombinations, and curators have used it chiefly to replace conventional record-keeping methods. The main thrust in record-keeping improvement has been to use the computer to emulate existing procedures, the benefits being the increased speed and accuracy. Little has been done so far to increase the scope of the data collected or the uses to which the data are put.

Data Conventions in Systematics

There are three basic levels of description in systematics. These are: Taxonomy - the application of names to plants and animals; Synonymy - the recording of both the taxonomic history of a species and the referrals to that (or other) species which can be verified by an author; and Morphologic description - the detailing of the features or characteristics of the species which discriminate it from others like it. These levels of description are developed further below, followed by a more detailed analysis of synonymy.

Taxonomy

About two hundred years ago Linnaeus imposed a degree of order on the naming of species of animals and plants by establishing the use of the generic-specific binomen. Each known plant and animal was given a generic name indicating its close genetic relationship to other species bearing the same name, and a species name to establish it as a unique and distinct entity within the genus. This formalized pattern of nomenclature replaced the use of vernacular names such as "horse" and "cow", provided a means by which naturalists could consistently describe new species for which no vernacular names existed, and provided a basis for hierarchical classification of animals and plants.

The naming of new species and the conservation of established specific names is very tightly controlled by the rules of biological nomenclature. The rules are universally recognized so that, in theory at least, all systematists should have reference to a single glossary of generic and specific names. Each name in the glossary should relate to one concept for a species. In turn, each species should be represented by one specimen or group of specimens originally designated by the describer as type specimens for that species. In practice the situation is not so clear cut. One problem is that it may be difficult or impossible for any systematist to keep up with the names in current use, even in his area of specialization, so that the same name may be used inadvertently and incorrectly for more than one species. Another problem is that the same species may be discovered and named independently by more than one researcher.

A different kind of problem arises when authors emend existing species concepts so as to more narrowly define an original fairly broad definition, or vice versa. While no name change may be involved, each re-definition changes the basis for reference of specimens to the species and considerable confusion can result. The problem is compounded when several revisers each apply their own re-definitions to an original species concept.

Synonymy

As part of his study of a plant or animal genus or species (taxon) an author will carefully consider identifications made of that taxon by other authors. He

will place in his synonymy for that taxon references to all previous citations which he believes belong to the taxon he is describing. Each such reference listed in a synonymy is a synonym. The basic information contained in each synonymic reference includes:

Generic name used by a previous author
Specific name used by a previous author
Name of previous author
Date of publication of the previous author's article
Name of the book or journal in which the article was published
Collation (pages, plates, figures) referencing the species in that article.

These items of information are shown in the first entry of the synonymy, depicted as Figure 1 (from Mello, 1971: 30) as follows:

Genus - Haplophragmoides
Species- excavata
Authors - Cushman and Waters
Date - 1927
Publication - Cushman Lab. Foram. Research Contr., V. 2, pt. 4
Pages, plates, figures - p. 82, pl. 10, figs. 3a, b

An author can use synonymic references to convey a great many different kinds of information or judgments about a species, or about the specimens which have been identified as belonging to a species. Some of these uses of synonymy, as shown on Figure 1, are described briefly below:

Heading - the genus and species names which head the synonymy are the current author's statement of what the described taxon should be called. The name(s) of the author(s) who first described the species in print is also given.

Affirmative references - citations in which the species has been described using the same specific name and concept as that of the current author. Complete synonymies will list all previous citations of this sort but it may be impractical to do this for species which have frequently been cited in the literature. The topmost synonym shown in Figure 1 is an example of an affirmative reference.

Reference recording taxonomic errors or differences in judgment - citations which place in synonymy taxa which previously had been recognized as separate. The second and third references in Figure 1 are examples of this type of change. Both Haplophragmoides glabra Cushman and Waters and Haplophragmoides kirki Wickenden are considered, in the paper cited, to belong to a single species with H. excavata. Placement of the original descriptions of these two taxa in synonymy with H. excavata has the effect of stating that the current author does not consider these taxa to exist.

Haplophragmoides excavata Cushman and Waters, emended

Haplophragmoides excavata Cushman and Waters, 1927, Cushman Lab. Foram. Research Contr., v. 2, pt. 4, p. 82, pl. 10, figs. 3a, b.

Haplophragmoides glabra Cushman and Waters, 1927, Cushman Lab. Foram. Research Contr., v. 2, pt. 4, p. 83, pl. 10, figs. 6a, b.

Haplophragmoides kirki Wickenden, 1932, Royal Soc. Canada Trans. 3d ser., v. 26, sec. 4, p. 85, pl. 1, figs. 1a-c.

North and Caldwell, 1964, Saskatchewan Research Council, Geology Div. Rept. 5, p. 13, pl. 1, fig. 8.

? Peterson, in Jones, 1953, Utah Geol. and Mineralog. Survey Bull. 47, p. 31, pl. 1, figs. 5a, b.

Haplophragmoides bonanzaensis Stelck and Wall. Tappan, 1962, U. S. Geol. Survey Prof. Paper 236-C, p. 133, pl. 30, figs. 16, 18 (not figs. 17, 19).

Mello, 1969, U. S. Geol. Survey Prof. Paper 611, p. 43, pl. 4, figs. 6a, b.

Figure 1

A Synonymy
(from Mello 1971: 30)

References to mis-identified specimens - citations which place into synonymy references to particular specimens cited in the literature under different taxonomic names. Examples of this kind of usage are the two references to Haplogragmoides bonanzaensis shown in Figure 1. The current author is stating that the attributions of specimens to H. bonanzaensis by Tappan in 1962 and Mello in 1969 are incorrect, and that both lots of specimens belong in H. excavata.

The examples described above and illustrated in Figure 1 are only some of the kinds of messages which can be embodied in synonymy, but they do provide an idea of the various roles which this information transfer medium is called upon to serve in systematics. There are many other constructions which can be used in synonymy to transmit different messages, including the use of question marks in various places as means of expressing degrees of uncertainty, and the use of parenthetical notes.

In principle, each synonymy should represent a particular author's judgment on the taxonomic limits of a species at the time he writes his paper. In principle, the specimens referred to in each citation included in the synonymy should have been examined personally by the author so that direct knowledge is used in placing them in synonymy with other specimens. In practice, it is often impossible for an author to personally examine all specimens which may belong in a species he is describing and thus he may have to rely on published illustrations and descriptions in deciding upon relationships. It is also unfortunately true that some authors uncritically accept the synonymies prepared by other authors and may transfer them bodily into their own papers.

In summary, there are many different kinds of relationships which can be expressed in a synonymy and there can be many degrees of reliability for the citations recorded. Nevertheless the composite of knowledge and interpretation embodied in the synonymies created for each species has considerable potential value to systematists. However, because of the volume of data and the complexity of interrelationships among data elements, it has so far not been possible to fully develop the information potential inherent in synonymy.

Morphologic Description

A third mode of expression for systematics data, much less rigorously structured than either taxonomy or synonymy, is found in morphologic description. There are three basic elements to morphologic description, namely - noun descriptors; adjectival and adverbial modifers; and syntax. By way of contrast, taxonomy consists only of noun descriptors (names), and synonymy consists almost exclusively of noun descriptors in a non-grammatical and quite rigid syntax. In morphologic description, as might be expected, there is much less control on the descriptive terms used.

For any particular group of organisms the noun descriptors which name the parts of the organism are usually well standardized and commonly understood. Thus all workers in foraminifera would have a common understanding of the term "suture" even though that term when used in the description of a trilobite would have a different connotation to specialists in that field. Adjectival and

adverbial modifiers, however, are almost always non- standardizeo as to their meaning. This common non-specificity of meaning for modifiers permits individual systematists to describe morphologic conditions or variations with a considerable degree of freedom. Consequently direct comparison between such descriptions is often difficult.

Use of the computer to systematically record morphologic descriptive information, at least for the foraminifera and a few other groups with which I am familiar, most likely must either await or accompany a fairly comprehensive standardization of the terms in which descriptions are written.

Computer Storage and Retrieval of the Data of Synonymy

The computer is a very literal machine. It cannot make extrapolations or interpretations. Thus any data to be compared by computer must be standardized in terms of information content and organization to the maximum degree possible. Synonymy is an information-rich mode of expression which also had the advantage of being largely standardized both in terms of format and in terms of data elements recorded. The computerization of synonymic data offers the benefit of creating a data base useful to systematists, and at the same time the complexity of possible data relationships provides an interesting test of computer capabilities.

Early in 1973, at a Museum Data Bank Coordinating Committee Meeting, Professor Jack Heller of the State University of New York - Stony Brook and I discussed the synonymy problem and agreed that from both the systematics and computer perspectives it offered interesting possibilities. Since that time we have been working on a method for the recording of synonymic data in machine readable form and testing the feasibility of the solutions arrived at. Neither of us have had more than a very limited time to work on the project and consequently we have not made great strides in terms of the volume of data stored. Nevertheless, we both feel that the hierarchical and linkage problems presented by synonymic data have been adequately faced and solved. The essentials of our work are outlined below.

Nature of the Synonymic Data Mode

Each newly identified species is validated by the author's selection of one specimen or a suite of specimens which are in effect the official name-bearers for the species. Once the concept of the species - represented by the range of variation described by the author and by the type specimens which he has designated - has been published and the species has received its official Linnean binomen, it can be referred to and cited by other workers. Many thousands of plant and animal species have already been described and form the base against which modern researchers compare their materials.

The synonymy created by an author-systematist for a taxon on which he is working is a succinct statement about previously described specimens which he feels belong in that taxon. It should be noted that the systematist must work

with the specimens, which form the objective basis for his efforts. Specimens are usually grouped into species on morphologic evidence, but sometimes genetic or behavioral evidence is used. In most cases it is impossible for the researcher to obtain independent evidence of the fact that two specimens belong in the same species (this can only be done through breeding experiments) and so usually the assignment of specimens to a species is a matter of judgment.

The first task of the systematist is to separate the specimens which he is working on into groups. For each group of specimens (presumed to be species) the researcher seeks to determine whether the specimens can be placed in any previously described species. Comparison of a group of specimens with type or accurately identified representative specimens of previously described species will often result in a positive match and assignment of the specimens to a previously named species. The comparisons the author has made and the conclusions he has drawn will be recorded in the synonymy.

In constructing a species synonymy an author will reference both original descriptions of species and later attributions. Inclusion in the synonymy of a citation to an original description equates the materials presently under study by the author with those used to describe the species originally. Inclusion of subsequent citations in the synonymy equates the specimens under study with those previously related either to the same taxon or to other taxa by other authors.

Analysis of Synonymic Data

The primary purpose to be served by computerizing synonymic data is to put the data of synonymy, and related information from the publication under analysis, in a form more directly useful to systematists. At the minimum, this information must be at least as accessible as it is in the conventional literature, and all existing relationships among the data elements must be preserved. Above the minimum level, computerization of synonymic data should make the elements of recorded information more accessible and more easily recombined than is the case in conventional synonymy. Optimally, computer storage of synonymic and related information should provide the means for a critical analysis of the adequacy and usefulness of synonymy as an information transfer tool, and might also provide insights for establishing new taxonomic and environmental relationships.

Synonymy provides a challenge from the point of view of computer data storage and retrieval partly because the most efficient means of recording the information for computer input does not necessarily organize the data in a fashion whereby it would be directly usable to systematists. For example, many taxa may have been recovered from a single sample taken at a discrete geographic location. All of the sample data should be related to each taxon, yet it makes no sense to enter the same data into the computer file for every taxon from the sample. In practice the information is gathered only once and linkages are established to each taxon by the computer program.

The data organization adopted for efficiency in data capture is shown in

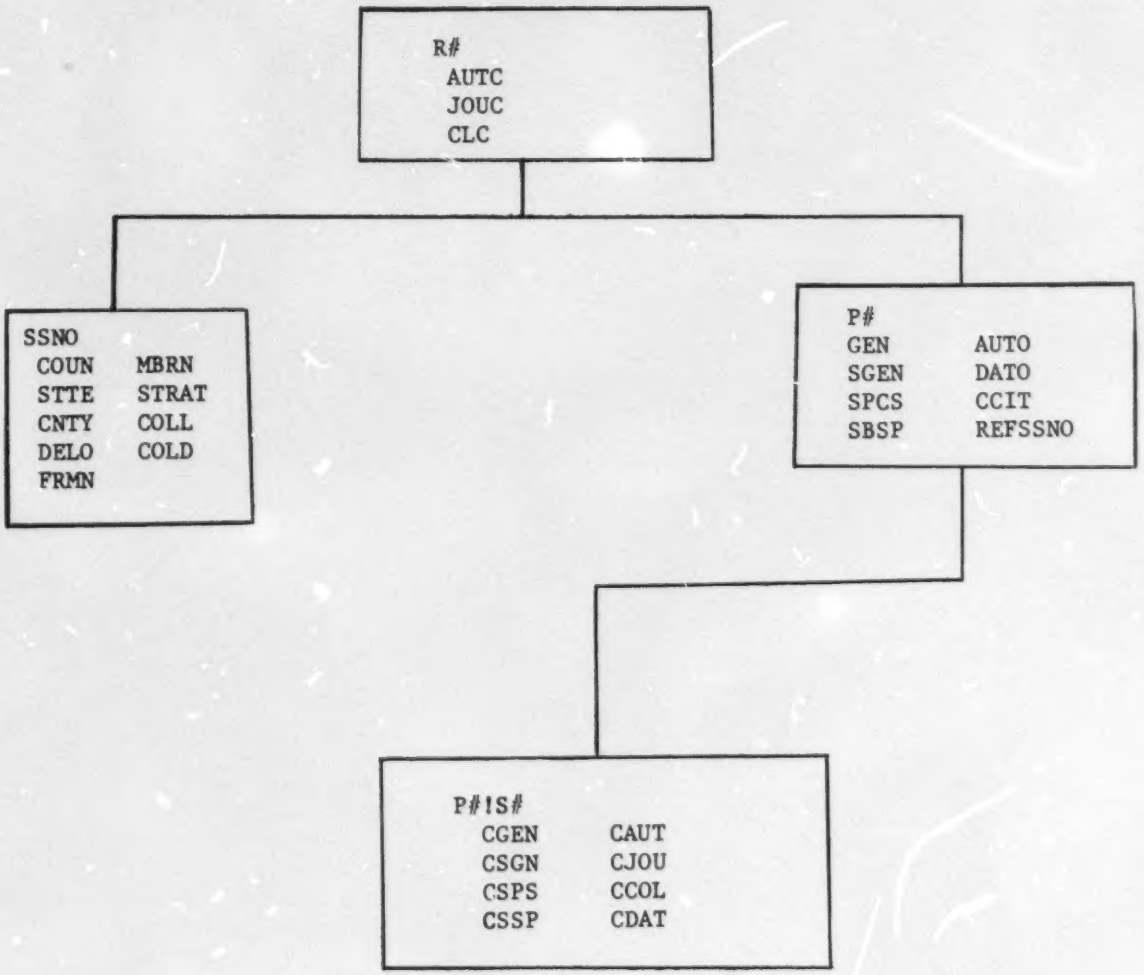

Figure 2

Synonymy data organization

Figure 2, and the data categories used are briefly defined as follows:

Explanation of Data Categories

R# - Identifying code for the article from which the data being recorded (current article) are derived.
AUTC - Name of the author (current author) of the current article.
JOUC - Journal name and volume in which the current article appeared.
CLC - pages, plates and figures composing the article.
SSNO - sample number, assigned from a single universe of sample numbers established for this study. Each sample collected by the author of the current article is given a number, and no two numbers are repeated in the file.
COUN - country name
STTE - state name
CNTY - county name
DELO - detailed locality information, such as township and range and distances from landmarks. This information should provide another collector the data he needs to find the sample locality.
FRMN - the name of the rock formation from which the sample was taken.
MBRN - the name of the member of the formation from which the sample was taken.
STRAT - the stratigraphic location for the sample, such as 50 feet above the base of the Carlile Shale, or 2 feet beneath the X bentonite.
COLL - name of the collector
COLD - date of collection.
P# - Each reference to a species which is entered into the file receives a number. That number designates the citation, whether it is listed in the synonymy or as the "primary" (P) citation in any given paper. A new species erected in a paper by Cushman in 1927 might be given the designation P274 when the Cushman paper is entered into the file. A subsequent citation of that identification, such as by Eicher in 1965, would appear in Eicher's synonymy and would then have the designation S274.
GEN - genus name used in a primary citation.
SGEN - subgenus name in a primary citation.
SPCS - species name in a primary citation.
SBSP - subspecies name in a primary citation.
AUTO - name of the original describer of the species. This name may be in parentheses, indicating that the species has been placed in a different genus than that to which it was originally ascribed.
DATO - date of publication of the paper in which the original description appears.
CCIT - the page and plate references to the description of this species in the current article. Not the page and plate data from the original reference.
REFSSNO - the number of a sample from which the species is reported in the current article. For each species all the samples in which it occurs are listed with REFSSNO tags. The sample numbers themselves are the same as those described under SSNO, but

different tags (SSNO and REFSSNO) re used to indicate the placement of the same data in two different parts of the data structure.

P#!S# - these are two numbers linked together to serve as a composite indicator for a reference which appears in synonymy. As discussed under the description for P#, each citation to a species is given a number when it is first entered into the file. The P or S prefixes are used to designate the reference as either "primary" or as synonym. The combination P#!S# used at the synonym level of the data hierarchy indicates that a certain citation number occurs as a synonym under a primary heading number. For example, the combination P274!S187 states that citation 187 occurs as a synonym under citation 274 in an article.

CGEN - the name of the genus cited in the synonymized citation.

GSGN - the subgenus name

CSPS - the species name

CSSP - the subspecies name

CAUT - the name of the author of the citation which has been synonymized. The name of the original describer of the species goes under AUTO.

CJOU - the name of the journal in which the article appears, including the volume number.

CCOL - the pages, plates and figures on which the cited taxon is described in the article.

CDAT - the date of publication of the cited article.

The data indicated in each box of Figure 2 are gathered at one time. Data sheets, shown as Figures 3-6, are used to record this information, with transcription to keypunch cards coming at a later stage. The first step, represented by the topmost box in Figure 2, is to record information about the current citation, using the data capture form depicted in Figure 3. The author of this work, the journal in which it appeared, the date of publication and the identifying number (R#) are recorded at this stage. Data related to each sample are recorded on the sample data sheet (Figure 4) and are shown in the left-most box of Figure 2. Information about each species described by the current author (right-most box of Figure 2) is recorded on the primary species data sheet (Figure 6) and is shown in the bottom box of Figure 2.

Linkages between these four convenient groupings of data are established by code number. The R# recorded with the current citation data is overlaid on all other records related to that citation by the computer program. A sample (SSNO) is applied by me to each discrete collecting locus reported by an author. The numbers are drawn from a single series beginning with "1" for the first sample from the first article entered and progressing from there. No sample number is assigned more than once. Sample numbers are cited in the primary taxon box after the REFSSNO tag. Each taxonomic citation, when it is first encountered, is given a unique number which will remain with it throughout the life of the file. Synonymic citations are linked to the primary citations which they are placed under through use of the "P" number of the primary taxon in conjunction with the "S" number of the citation itself.

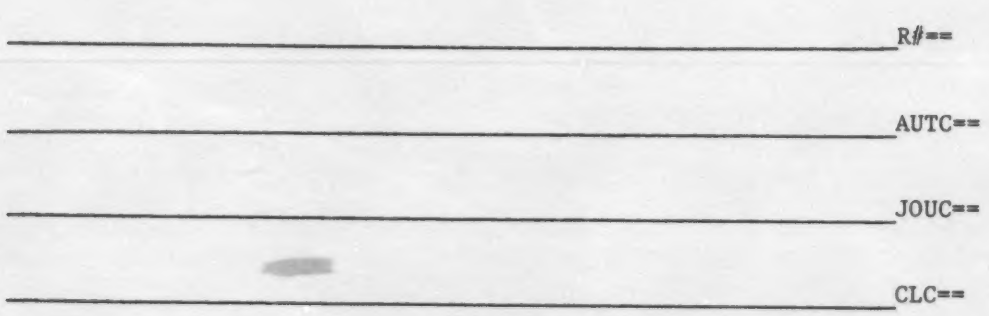

Figure 3

Current citation data capture form

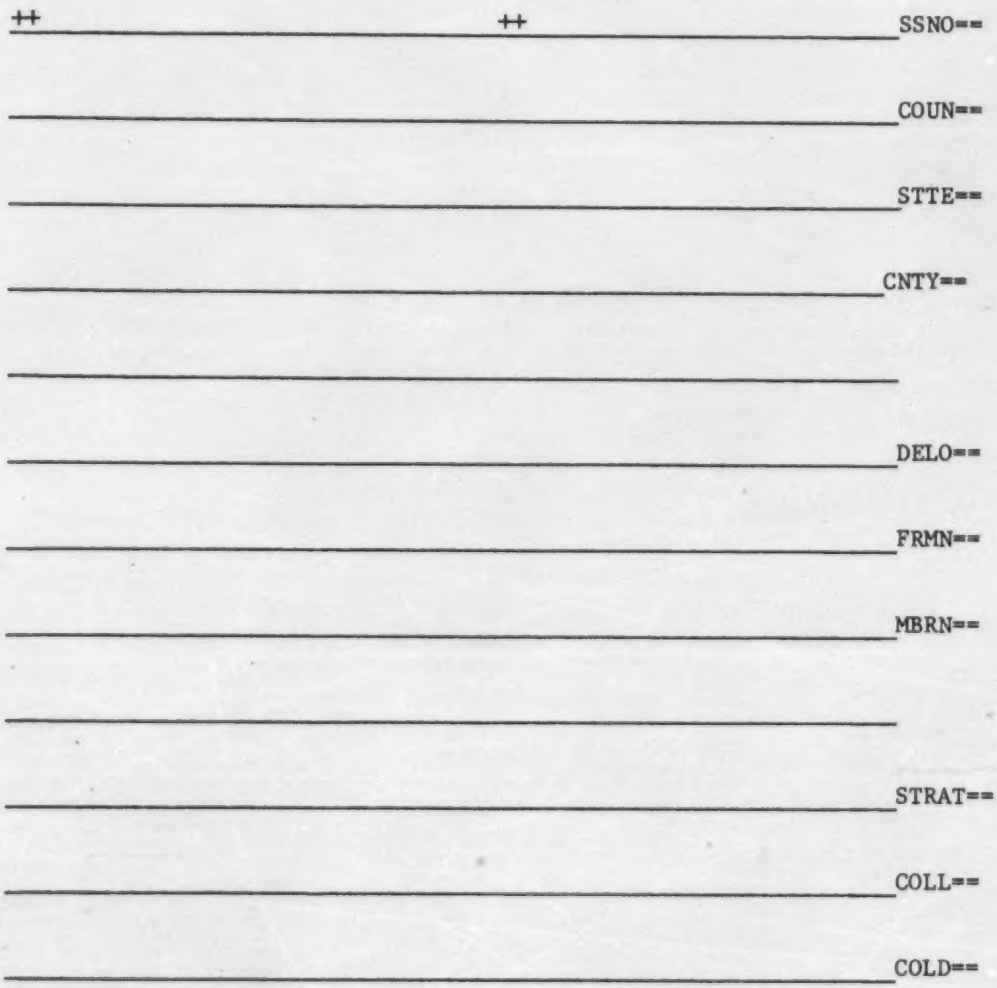

Figure 4

Sample data capture form

++P _____ ++

_____GEN==

_____SGEN==

_____SPCS==

_____SBSP==

_____AUTO==

_____DATO==

_____CCIT==

_____REFSSNO== _____REFSSNO==

_____REFSSNO== _____REFSSNO==

_____REFSSNO== _____REFSSNO==

_____REFSSNO== _____REFSSNO==

_____REFSSNO== _____REFSSNO==

_____REFSSNO== _____REFSSNO==

_____REFSSNO== _____REFSSNO==

Figure 5

Species data capture form

++P _____ !S _____ ++

_____CGEN==

_____CSGN==

_____CSPS==

_____CSSP==

_____CAUT==

_____CJOU==

_____CCOL==

_____CDAT==

Figure 6

Synonymy data capture form

Query Capabilities

The labor of gathering the citation, sample, primary taxonomic and synonymic data described above, is repaid in the form of querying capability. The Griphos computer programs provide the means to link the data elements in many useful ways.

For generations botanists and zoologists have been recording and cross-referencing information about their taxa. The conventional method has been to create one or a few 3 x 5 card files - certainly an alphabetic listing by genus and species, perhaps a cross-reference file to the localities from which specimens were obtained, and undoubtedly a bibliography of papers dealing with the subject matter. Files such as these represent a substantial investment of both scholarship and labor and usually are individually designed and unique. Upon the death or retirement of the specialist who creates such a file it is all too often the case that the file is not maintained and soon becomes useless. Even if it is maintained, however, it almost always exists as one discrete entity which cannot easily be duplicated for use at other locations.

The purpose of the work described here is to make it possible to efficiently capture and flexibly organize date to query those data in ways which are not possible with card files, and to duplicate easily the file for the benefit of others who need it.

Some of the interesting relationships which can be established from these data through the use of the Griphos programs are described below:

Synonyms in Common

With the passage of time each published taxonomic citation is likely to find its way into the synonymies of later authors. The placement of earlier citations into a synonymy by an author not only establishes a relationship between the citations which appear together in the synonymy, but also allows for the accumulation and listing of all of these co-occurrences as a useful guide to taxonomic thinking at various times in the past as well as an indication of the likely relationships of taxa within the group.

Sample - Taxa Relationships

As a matter of convention most papers which describe fauna or flora of a region order their taxa in terms of the inferred genetic relationships between them. Thus the relationships which exist between the taxa and the samples from which they came is often obscured. The computer provides the means to recombine taxa and sample information. Results might include listings of taxa from particular rock units (for fossils) or sediment types, or maps of geographic distribution by State or County.

Co-occurrence of Species

In sampling, whether for fossil or recent organisms, it is often found that many different species exist in one spot. Often the natural

interrelationships of co-occurring species are lost or obscured in the article in which they are described. The computer can re-establish these relationships and produce lists of co-occurring species which could well be of value to future workers on the group.

Results

Work done so far includes the entry of data on three papers dealing with microscopic fossils from the western United States, and work with the Griphos system to adjust to the particular complexities of the hierarchical data being treated. While the data base itself is still too small to be interesting for scientific research purposed, the results of experiments in recombining the data have been most promising. Some of the products are described below:

Frequency dictionaries -

These are listings, ordered alphabetically or numerically, of each data element in the file. Each new data element is listed once and all references to it (in the computer master file) are indicated by record number.

Genus Name	Record Number
Anomalina	115, 205, 222
Bigenerina	27, 118
Globigerina	10, 245

Each genus name in the example above is listed just once and all records in the file where reference to it can be found are listed to the right. The same kind of list is made for species names, journal citaions, locality data and so forth. Such lists are valuable because they greatly facilitate editing of the files - since each term is listed once and any spelling or punctuation variation is recognized as a separate term by the computer and listed separately - and as ready references to what is contained in the file.

Taxonomic lists -

Various kinds of taxonomic lists are potentially useful, such as those described below:

Genus	Species	Author	Date
Anomalina	henbesti	Morrow	1936
Bigenerina	bidentata	Loetterle	1937
Bigenerina	maculata	Eicher	1965

Genus	Species	Author	Date
Cited Genus	Cited Species	Cited Author	Cited Date
Anomolina	henbesti	Cushman	1931
Anomolina	henbesti	Cushman	1932
Anomolina	infrequens	Jones	1930
Gyroidina	depressa	Plummer	1928
Gyroidinoides	depressa	Chapman	1931

Sample lists -

Data derived from samples can be recombined in various ways of interest to researchers, such as in the examples which follow:

Formation	Member	State	County	Sample Number
Carlile Shale		Colorado	Freehold	11
Carlile Shale	Frontier	Colorado	Denver	27
Pierre Shale	Mobridge	South Dakota	Corson	175
Pierre Shale	Verendrye	South Dakota	Dewey	180

State	County	Formation	Member	Sample Number
South Dakota	Corson	Pierre Shale	Mobridge	175
South Dakota	Corson	Pierre Shale	Mobridge	179
South Dakota	Dewey	Pierre Shale	Verendrye	180
Wyoming	Custer	Carlile Shale	Frontier	77

Selective lists -

It is possible to first select and then list data from a large file to satisfy particular needs. For example, a listing of all species from the Carlile Shale of Wyoming could be made by first selecting or sorting out the records pertaining to those two data elements and then listing those records alphabetically by genus and species. There are many other combinatorial possibilities which could also be of interest to researchers.

Conclusion

The taxonomic, synonymic and geographic data normally recorded in biological systematics are well standardized and thus lend themselves to computerization. Yet the relationships among the data elements are complex. This study has concentrated on developing efficient methods of data entry and computer processing. In the future, when the file contains a significant proportion of the information available about the fossil group and geographic region I am dealing with, it should become increasingly useful as a source of information on research questions.

REFERENCES CITED

Mello, James F.

1971 Foraminifera from the Pierre Shale (Upper Cretaceous) at Red Bird, Wyoming. U.S. Geological Survey Prof. Paper 393-C: 1-54. pls. 1-7.